IN THE UNITED STATES
PATENT AND TRADEMARK OFFICE

Patent Application

Inventor(s)	Mark Beutnagel Ariel Fischer Joern Ostermann Yao Wang	Case Name	Beutnagel 4-1-13-3
Filing Date	12/31/1998	Serial No.	09/224,583
Examiner	Michael Opsasnick	Group Art Unit	2654
Title	Integration of Talking Heads and Text-to-Speech Synthesizers for Visual TTS		

ASSISTANT COMMISSIONER FOR PATENTS
WASHINGTON, D.C. 20231
SIR:

DECLARATION UNDER 37 CFR 1.132

1. My name is Hans Peter Graf.
2. I have a PhD in Physics from the ETH in Zurich, Switzerland, awarded to me in 1981.
3. I have been employed at AT&T since 1983, and have been involved in image generation and analysis for 15 years.
4. I am the holder of over a dozen US patents.
5. I was requested by Mr. Beutnagel, one of the inventors in the above-identified application to read US Patent 5,884,267, issued to Goldenthal, and to state my understanding of what it teaches, and what it does not teach (that is relevant to the above-identified application).
6. Goldenthal describes a system and a method for creating computer generated images and associated audio from a received audio signal.
7. A speech signal is acquired by microphone 110 and is converted to digital form in A/D converter 120. The digital signal is then applied to lines 111 and 112.
8. Via line 111, the digital signal is applied to a DSP 200, which converts the digital signal to acoustic-phonetic units, i.e., phonemes.
9. A translator unit 130 converts the acoustic-phonetic units to visemes, on line 116.

10. Thus, the entire path between microphone 110 and line 116 can be viewed as a single unit that converts in incoming audio signal to visemes, based on a dictionary that converts phonemes to visemes. Each viseme has a starting time and an ending time.
11. The digital signal of line 112 is placed in an audio file having, for example, a ".wav" format.
12. The audio file and the visemes are applied to rendering element 240.
13. There is no teaching whatsoever relative to the processing within rendering element 240, other than a reference to a patent application bearing the Serial No. 08/258,145.
14. It is quite clear to me that the Goldenthal reference does not deal with TTS signals. It deals only with audio signals (either from a microphone, or from a ".wav" file transmitted over the Internet).
15. It is also quite clear to me that the Goldenthal reference does not deal with FAP signals (markers, bookmarks) that are explicitly embedded in the input signal.
16. The Goldenthal arrangement cannot process facial expression information at an input signal, if that were provided.
17. Consequently, the image created by the Goldenthal reference cannot have facial expressions, such as head nods, eye movements, etc.
18. Even if the rendering system 240 of Goldenthal were to be designed to provide some facial expression (for which there is no teaching whatsoever, or any suggestion) still it is a fact that there is nothing in the input that specifies such facial expressions.
19. I also read US Patent 4,884,972 (Gasper). It describes a system where a user is presented with tile images. Each tile has an imprinted image of a letter of a phonogram.
20. The user moves tiles, and when concatenated tiles form words, a sound is formed corresponding to the concatenated tiles.
21. While the sound is made, a talking head image is changed in synchronism.
22. It provides no input that corresponds to a specification of facial expressions.
23. Consequently, combining the Goldenthal referenced with the Gasper reference does not result in an arrangement where facial expressions are specified by the input, or an arrangement that can accommodate (i.e. process) facial expressions that are specified in the input signal.
24. Finally, I also read US Patent 6,130,679 (Chen et al).

25. This reference develops some (i.e., a small amount of) FAP information from an input video. It does not create facial expression information from the audio. Therefore, it cannot synchronize the synthesized images to audio.
26. Chen's description basically focuses on how to reduce the amount of data that can be obtained from an video signal so that a minimum amount of data can be transmitted, and interpolated on the receiving end.

Respectfully,

Dated: 10-14-2002